Implications of encode on cancer biology essay

Science, Biology

Human Genome Project (HGP): The HGP was initiated in 1990 with the objective of determining the DNA sequence of the entire human genome within 15 years. By 2001, there was a draft sequence of the human genome4. Researchers identified, in its 3 billion letters, many of the regions that code for proteins. These protein regions, or exons, make up little more than 1% of the genome and contain around 20, 000 genes4. One particularly striking finding of the HGP research is that the human nucleotide sequence is nearly identical (99. 9%) between any two individuals. Moreover, a single nucleotide modification in a gene can be enough for causing human diseases such as cancer5. But almost 90% of these variants occur outside of protein-coding genes, leaving researchers with little information as to how they might cause or influence disease6. Thus, a deeper understanding of the human genome sequence was neededto gain a further understanding of the molecular mechanisms underlying a multitude of human diseases, including cancer. ENCODE Project: The function of the vast majority of the human genome has remained largely unknown, but the Encyclopedia of DNA Elements (ENCODE) project, launched in 2003, set out to change that7 (Fig. 1). ENCODE was designed to pick up where the HGP left off and is the most significant advance made in not only cancer research, but all medical research, in the last year. No Junk DNA: Although the HGPs massive effort revealed the blueprint of human biology, it quickly became clear that there was a lot more to the genome than anyone could have predicted. ENCODE revealed that the human genome is more 'functional' than researchers had believed. Although less than 2% of the genome codes for actual proteins, the ENCODE project indicated that about 80% of the genome is active, with the

largest class representing the different RNA types, covering 62% of the genome6, 8. ENCODEs effort has revealed that a gene's regulation is far more complex than previously thought, being influenced by multiple stretches of regulatory DNA and by strands of RNA not translated into proteins. Now no matter which part of the genome researchers are studying, they will benefit from looking up the corresponding ENCODE data and filling in the functional data. Non-Coding But Functional: One objective of the ENCODE project was to identify all the RNA molecules transcribed from DNA and to determine all the transcription initiation sites9. Djebali et al. report evidence that three-quarters of the human genome is capable of being transcribed, as well as observations about the range and levels of expression, localization, processing, and modifications of almost all known and thousands of new RNAs10. For example, ENCODE defined 8800 small RNA molecules and 9600 long non-coding RNA (lncRNA) molecules10. LncRNAs have been found to be deregulated in several human cancers and show tissue-specific expressio11. One example previously found, that was also sequenced by ENCODE, was the lncRNAs HOTAIR, which can cause changes to the chromatin state and promote metastasis12, 13 (Fig. 2). This is important because it indicates that nearly all of our genome is dynamic and active. As well, these results from the ENCODE project show that most stretches of DNA are active in regulating the expression of genes14. This should stimulate new research directions and therapeutic options considering non-coding RNAs as novel markers and therapeutic targets. Chromatin Modifications: Another interesting finding of the ENCODE project is the unraveling of the transcriptome and how epigenetics can regulate

transcription. It is now known that ~75% of our genome is transcribed10. One ENCODE research team looked at the relation between open chromatin configuration (euchromatin) and gene expression. They found a clear relationship between euchromatin and high gene expression, demonstrating that chromatin configuration is used to regulate gene expression (so much so that gene expression can be predicted from chromatin structure)15 (Fig. 3). Furthermore, the project looked into how chromatin structure correlated with transcription factor binding sites to show that down-regulated genes often had their transcription factor binding-sites obscured via tightly-packed (heterochromatin) chromatin16. These results have given a clear, defined map of transcription and also a clear idea of how transcription is regulated epigenetically. ENCODE will help unravel the cause and effect relationship between the epigenetic modifications and their contribution to diseases like cancer, and help determine if certain epigenetic marks can also be used as therapeutic targets. Long Range Signals: The vast non-coding portion of the human genome is full of functional elements and disease-causing regulatory variants. Spatial proximity and specific long-range interactions between genomic elements can be detected using a technique called chromosome conformation capture (5C) (Fig. 4). Using this technique, researchers looked for places where DNA from distant regions of a chromosome, or even different chromosomes, interacted17. They found that an average of 3. 9 distal stretches of DNA linked up with the beginning of each gene. As well, in each cell line they discovered <1, 000 long-range interactions between promoters and distal sites that include elements resembling enhancers, promoters and CTCF-bound sites10, 18. Additionally, they identified large

numbers of transcriptional start site-distal fragment interactions, of which ~60% were observed in only one of the three cell lines14 (Fig. 5). These data point to intricate cell-type-specific 3D folding of chromatin. Furthermore, these results have started to place genes and regulatory elements in 3D context, revealing their functional relationships. With increases in sequencing capacity, similar high-resolution studies should become feasible to map specific long-range interactions throughout the genome, which may uncover further principles that guide chromatin and gene interaction. Such insights will also be critical for interpreting genome-wide association studies (GWAS) that often identify regions with regulatory elements but not their distally located target genes. Assigning Function to GWAS Data: Since 2005, GWASs have identified thousands of points on the genome in which a single nucleotide variant seems to be associated with disease risk19. But almost 90% of these variants fall outside of protein-coding genes, so researchers have little clue as to how they might cause or influence disease6. The information provided by the ENCODE project now makes it possible to assign a function to many of the single nucleotide polymorph isms (SNPs) identified in non protein-coding regions of the genome which will surely have a large impact on the prognosis and diagnosis of these diseases, and might even open the way to new therapeutic strategies. Schaub et al. investigated multiple types of functional data generated by the ENCODE Consortium to help identify 'functional' SNPs that may be associated with the disease phenotype. Their work shows putative function for up to 80% of all previously reported associations20. They were able to confirm two previously proposed prostate cancer markers, rs902774 and rs229884, using the ENCODE

data14. ENCODE data has also shown that the majority of phenotype-associated SNPs in the GWAS catalogue are enriched in nucleosomefree regions bound by transcription factors21. Their results show that the experimental data sets generated by the ENCODE project can be successfully used to suggest functional hypotheses for variants associated with cancer. Thus, high throughput genomic assays are providing significant aid to our understanding of how SNPs, identified by GWAS, can contribute to human cancers. Implications of ENCODE on cancer: The ENCODE project aimed to fully describe the list of functional elements that make up the genome. The various groups report that the non-protein coding regions of the genome are filled with enhancers (regulatory DNA elements), promoters (the sites at which the transcription of DNA into RNA is initiated) and various regions that encode RNA transcripts that are not translated into proteins but have regulatory roles. Until now, the focus had largely been on looking for errors within genes themselves, but the ENCODE research will help guide the hunt for problem areas that lie elsewhere in our DNA sequence. Recent evidence points to 40% -60% cases of many common tumors have at least one mutation that might impact therapeutic decision-making or might suggest enrollment in a certain clinical trial22, 23. Therefore, the ultimate test of cancer genomics will be its ability to improve diagnostics and therapeutics. As sequencing costs fall, diagnostics may move to whole-genome sequencing. The challenge will be to decipher the data for clinicians to allow better treatment for the patients. ENCODE is a step towards a more complete understanding of the entire human genome and eventually,

genomic analysis will likely become part of the standard of care for cancer patients and ___